M2 BIM/STRUCT - Lecture 3

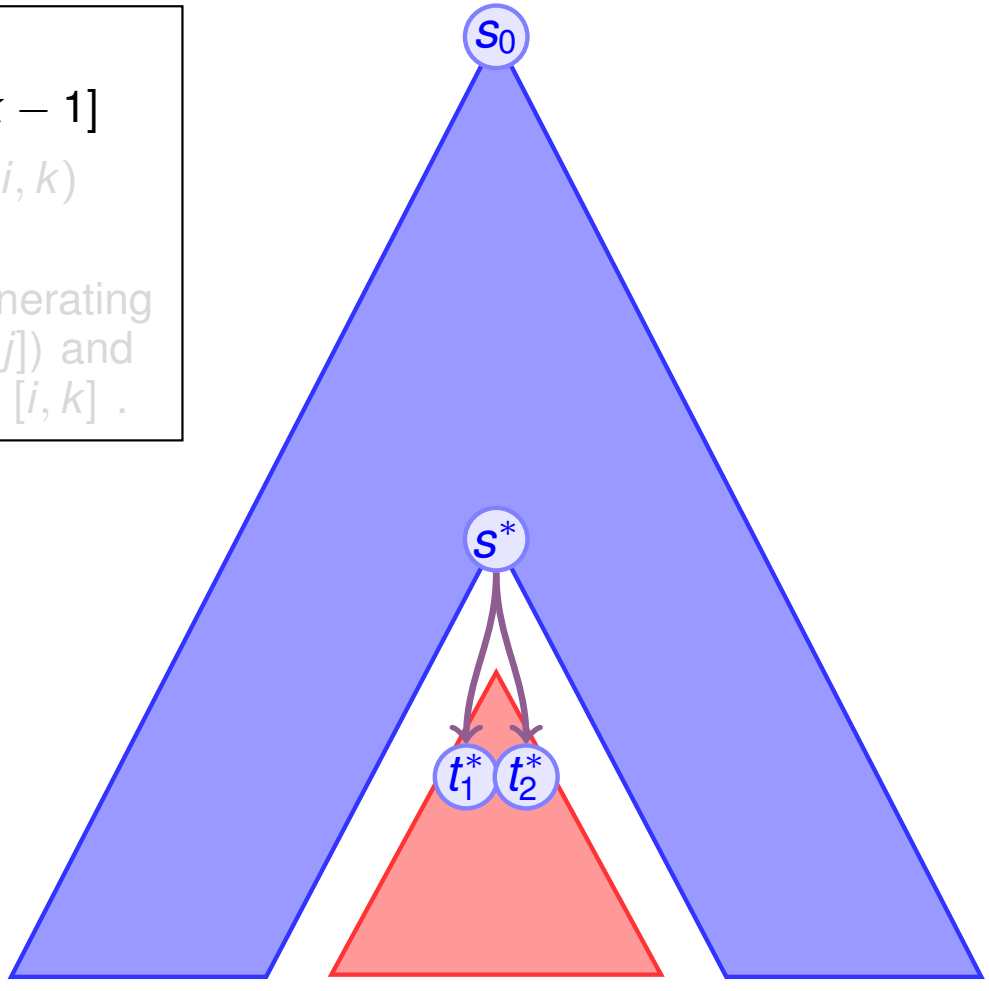
Advanced dynamic programming and alignment

Yann Ponty

AMIBio Team
École Polytechnique/CNRS

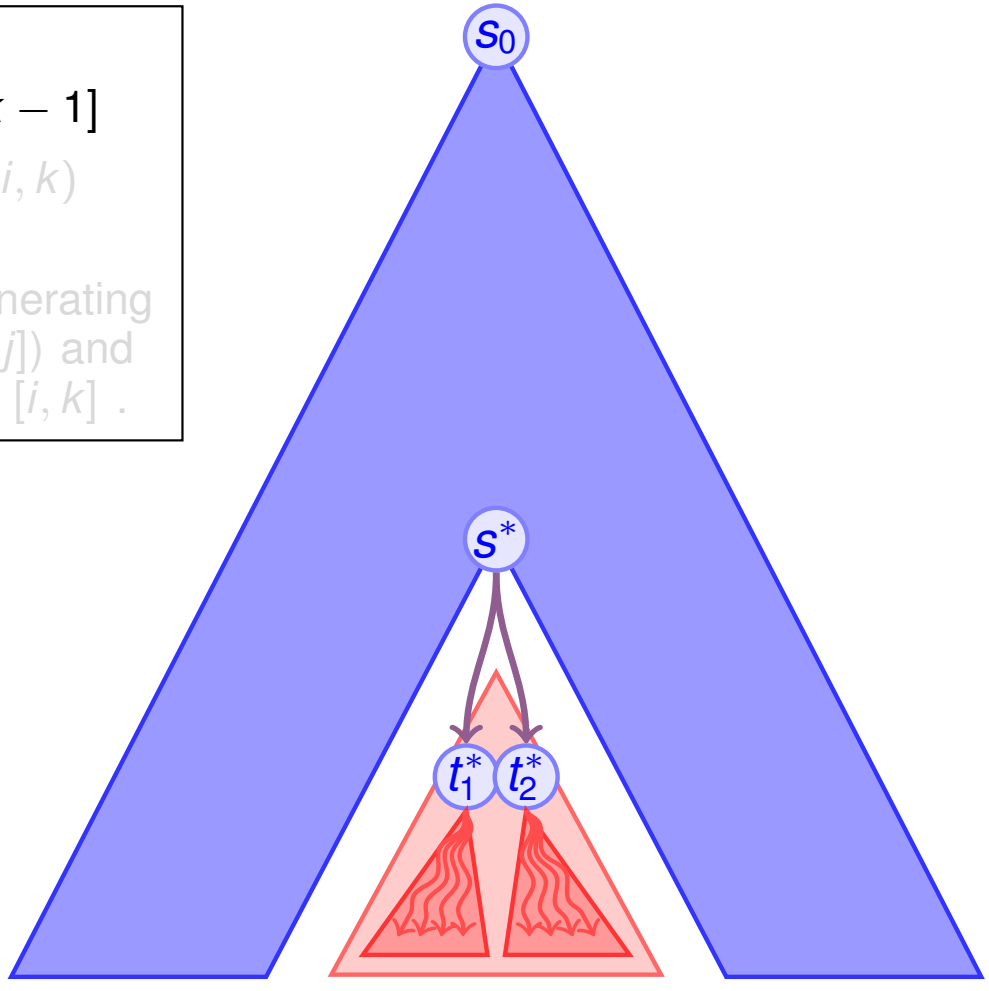
Structure including base pair (i, k):

- Inside: Structures over $[i + 1, k - 1]$
- Outside: **Contexts** of interval (i, k)
  - $\forall$ interval $[i, j], i < j \leq k$
  - **Complete** structure by generating **brother intervals** ($[k + 1, j]$) and recurse over the **father** of $[i, k]$.



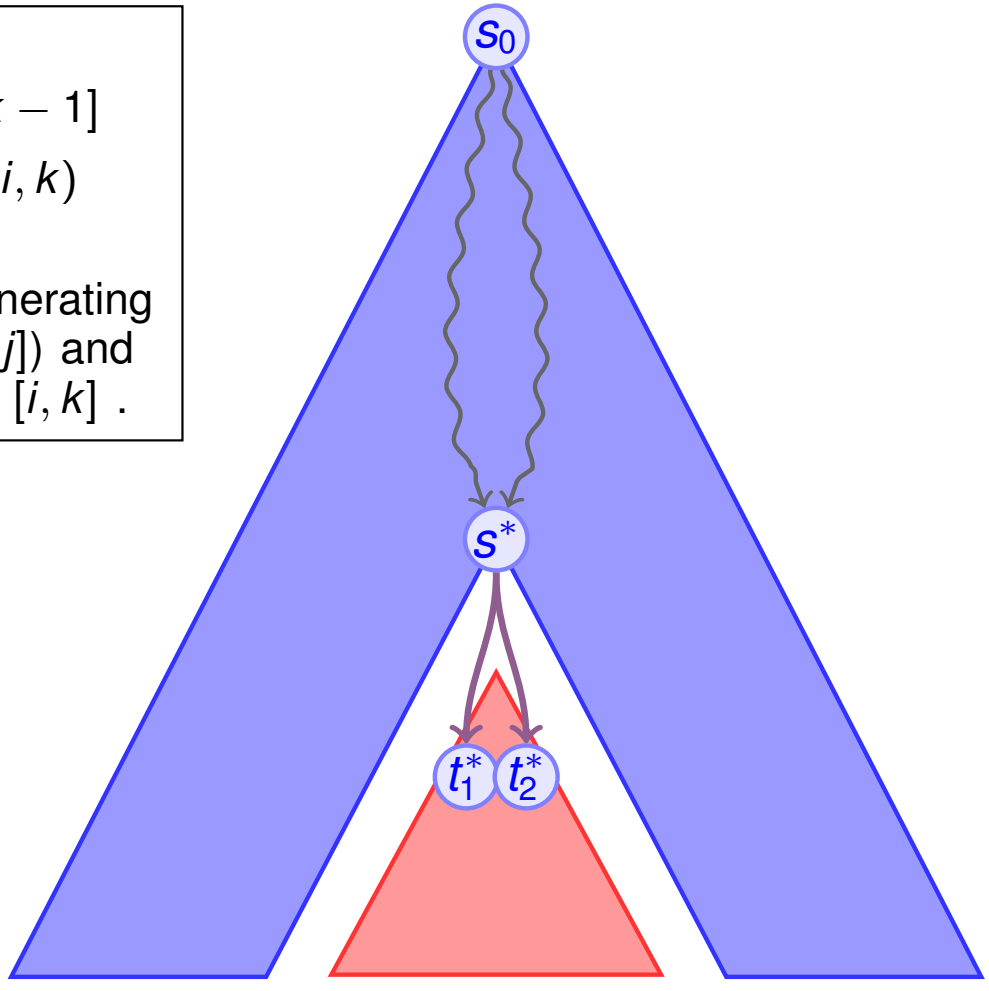
Structure including base pair (i, k):

- Inside: Structures over $[i+1, k-1]$
- Outside: **Contexts** of interval (i, k)
  - $\forall$ interval $[i, j], i < j \leq k$
  - **Complete** structure by generating **brother intervals** ($[k+1, j]$) and recurse over the **father** of $[i, k]$.



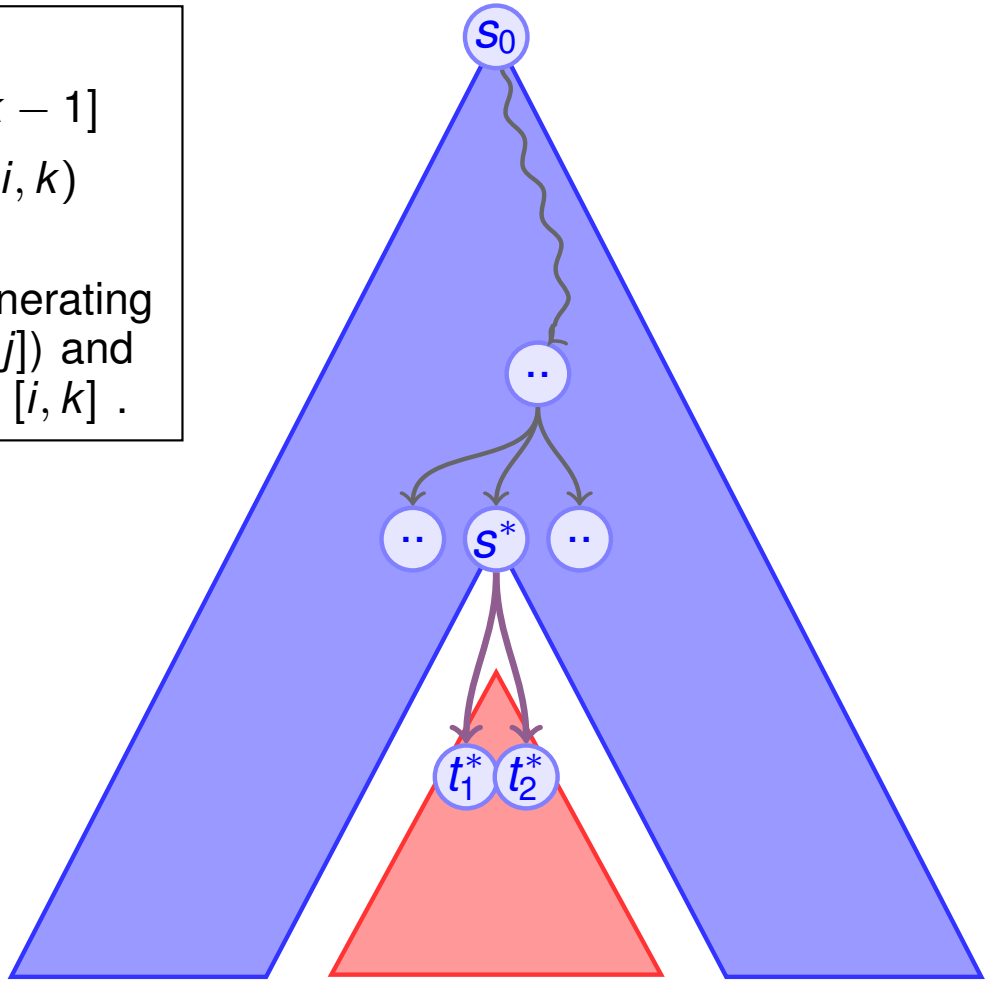
Structure including base pair (i, k):

- Inside: Structures over $[i+1, k-1]$
- Outside: **Contexts** of interval (i, k)
  - $\forall$ interval $[i, j], i < j \leq k$
  - **Complete** structure by generating **brother intervals** ($[k+1, j]$) and recurse over the **father** of $[i, k]$.



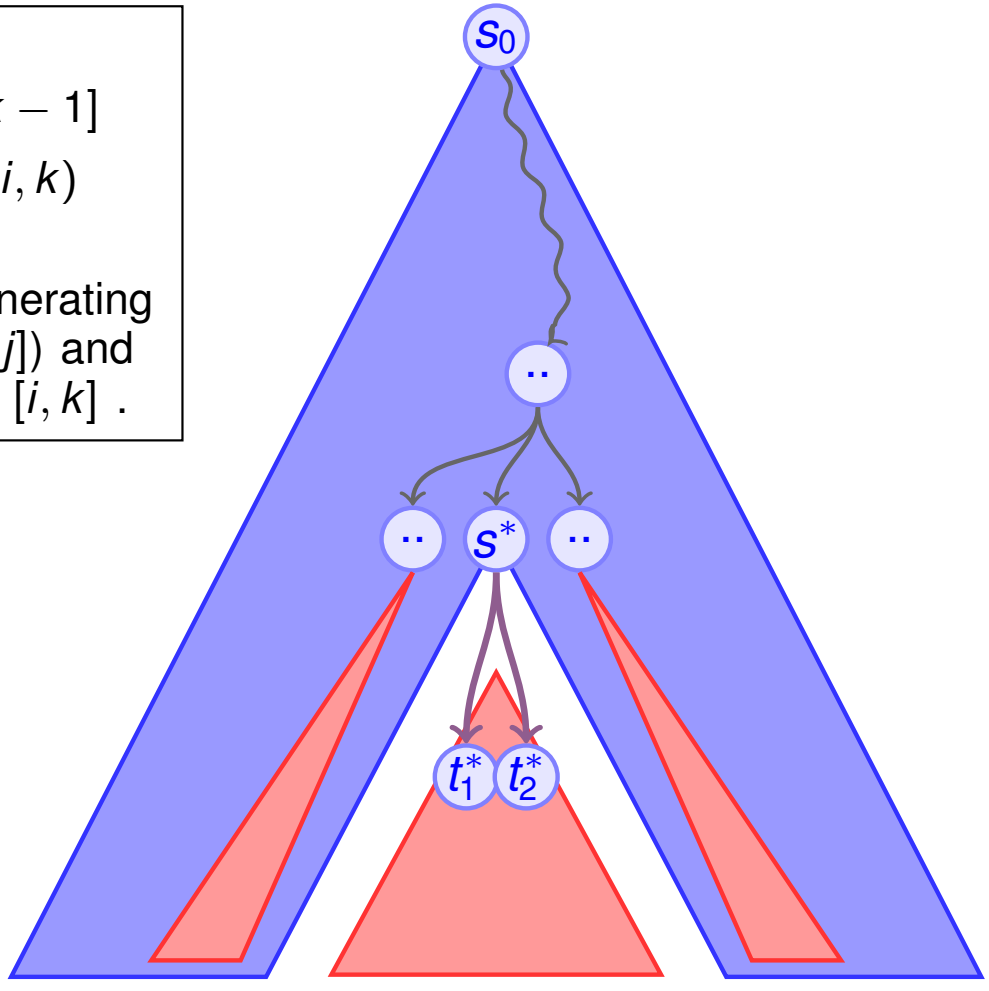
Structure including base pair (i, k):

- Inside: Structures over $[i+1, k-1]$
- Outside: **Contexts** of interval (i, k)
  - $\forall$ interval $[i, j], i < j \leq k$
  - **Complete** structure by generating **brother intervals** ($[k+1, j]$) and recurse over the **father** of $[i, k]$.



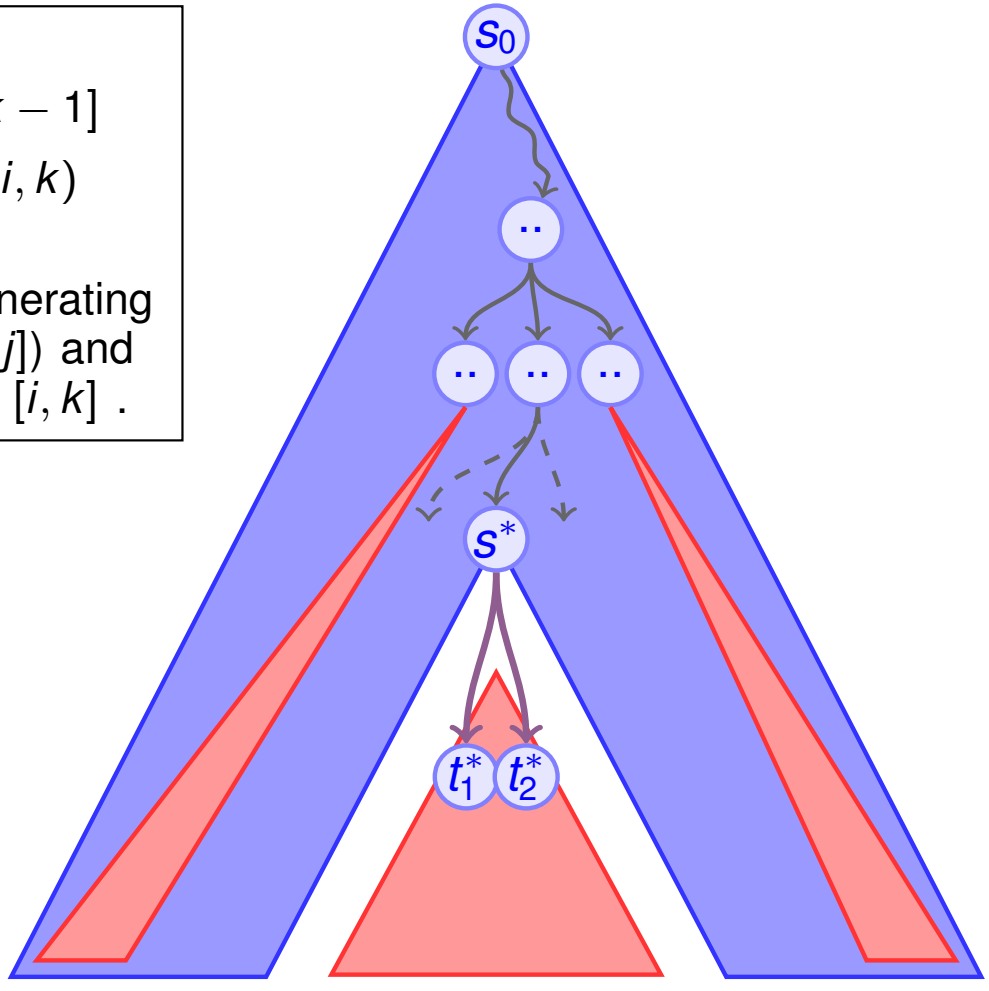
Structure including base pair (i, k):

- Inside: Structures over $[i+1, k-1]$
- Outside: **Contexts** of interval (i, k)
  - $\forall$ interval $[i, j], i < j \leq k$
  - **Complete** structure by generating **brother intervals** ($[k+1, j]$) and recurse over the **father** of $[i, k]$.



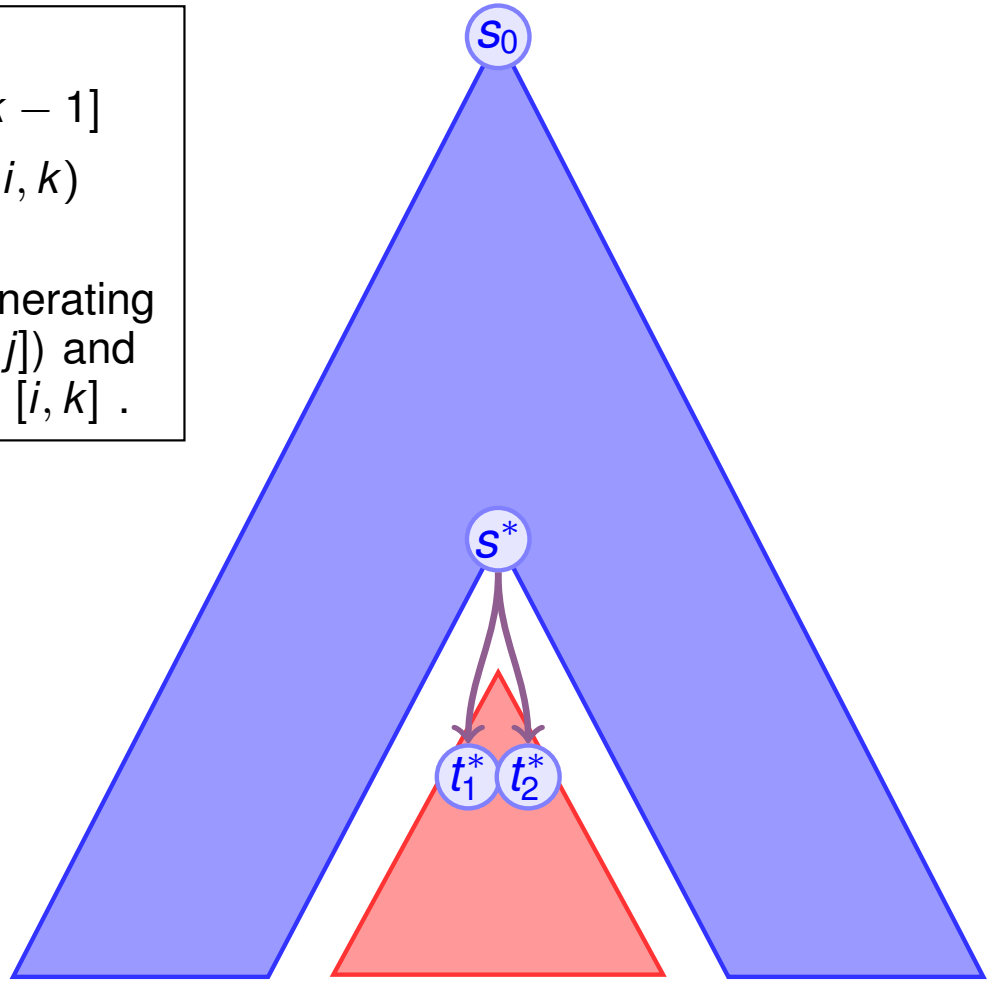
Structure including base pair (i, k):

- Inside: Structures over $[i+1, k-1]$
- Outside: **Contexts** of interval (i, k)
  - $\forall$ interval $[i, j], i < j \leq k$
  - **Complete** structure by generating **brother intervals** ($[k+1, j]$) and recurse over the **father** of $[i, k]$.



Structure including base pair (i, k):

- Inside: Structures over $[i+1, k-1]$
- Outside: **Contexts** of interval (i, k)
  - $\forall$ interval $[i, j], i < j \leq k$
  - **Complete** structure by generating **brother intervals** ($[k+1, j]$) and recurse over the **father** of $[i, k]$.



Whenever some further **technical conditions** are satisfied, this decomposition is **complete** and **unambiguous**, and implies a ***simple recurrence*** for computing the base pair probability matrix in $\Theta(n^3)$.

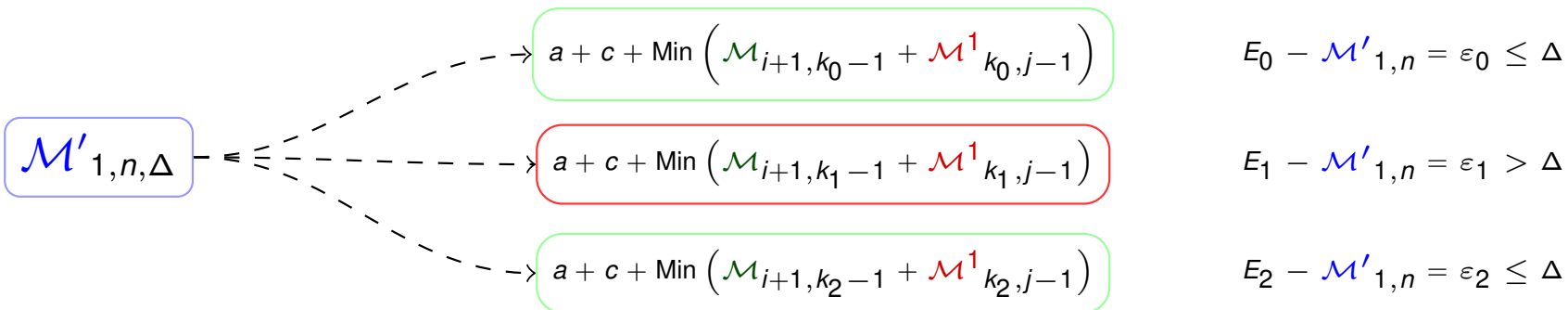
Alternatively: Duplicate sequence

$\rightarrow$ Inside contribution over $[j, n] \cup [1^\star, j^\star]$ = Outside contribution of $[i, j]$.

Prob.: Simplified energy model (no pseudoknots, only canonical BPs)
$\Rightarrow$ **Native** structure (functional) could be **overthrown**.

$\Rightarrow$ Investigate suboptimal structures (`RNASubopt` [WFHS99]),
i.e. build all structures within Δ KCal.mol^{-1} of MFE:

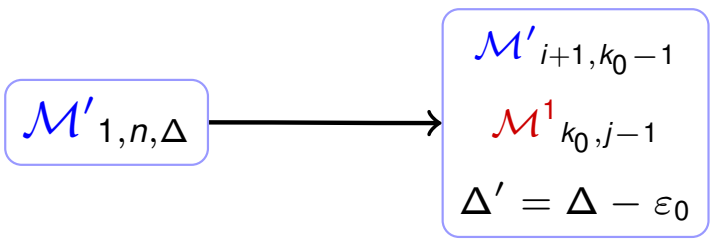
- Compute minimum free-energy matrices
- Backtrack on any contribution within Δ of MFE;
- Update Δ such that future backtracks create ≥ 1 struct.
- Recursively generate subopts and combine (brute-force ou Sort)



Prob.: Simplified energy model (no pseudoknots, only canonical BPs)
$\Rightarrow$ **Native** structure (functional) could be **overthrown**.

$\Rightarrow$ Investigate suboptimal structures (`RNASubopt` [WFHS99]),
i.e. build all structures within Δ KCal.mol^{-1} of MFE:

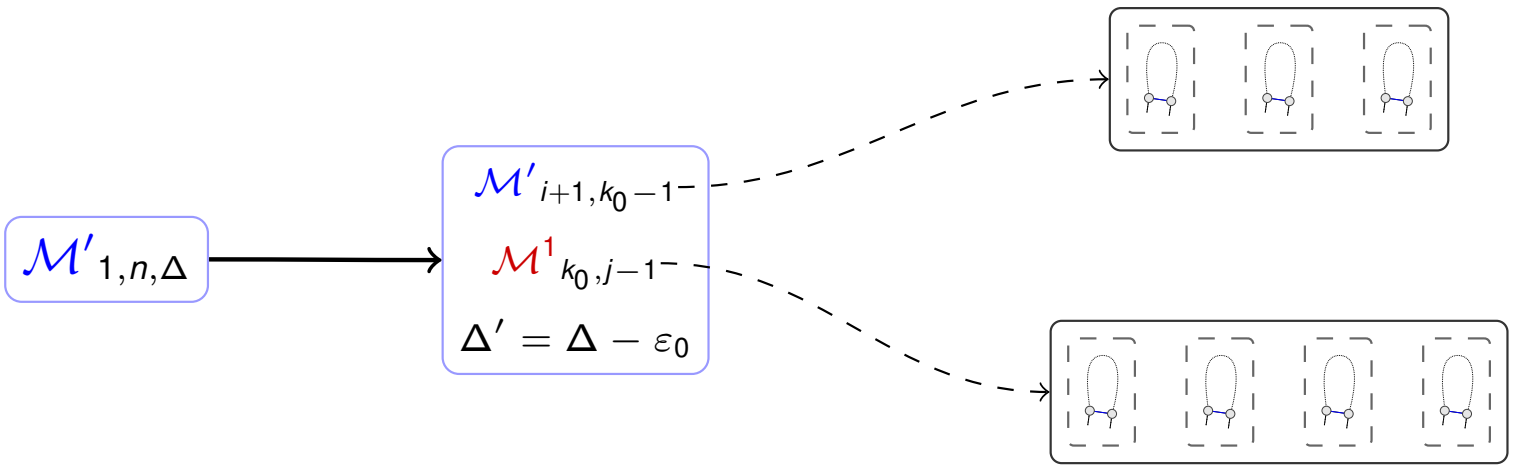
- Compute minimum free-energy matrices
- Backtrack on any contribution within Δ of MFE;
- Update Δ such that future backtracks create ≥ 1 struct.
- Recursively generate subopts and combine (brute-force ou Sort)



Prob.: Simplified energy model (no pseudoknots, only canonical BPs)
$\Rightarrow$ **Native** structure (functional) could be **overthrown**.

$\Rightarrow$ Investigate suboptimal structures (`RNASubopt` [WFHS99]),
i.e. build all structures within Δ KCal.mol^{-1} of MFE:

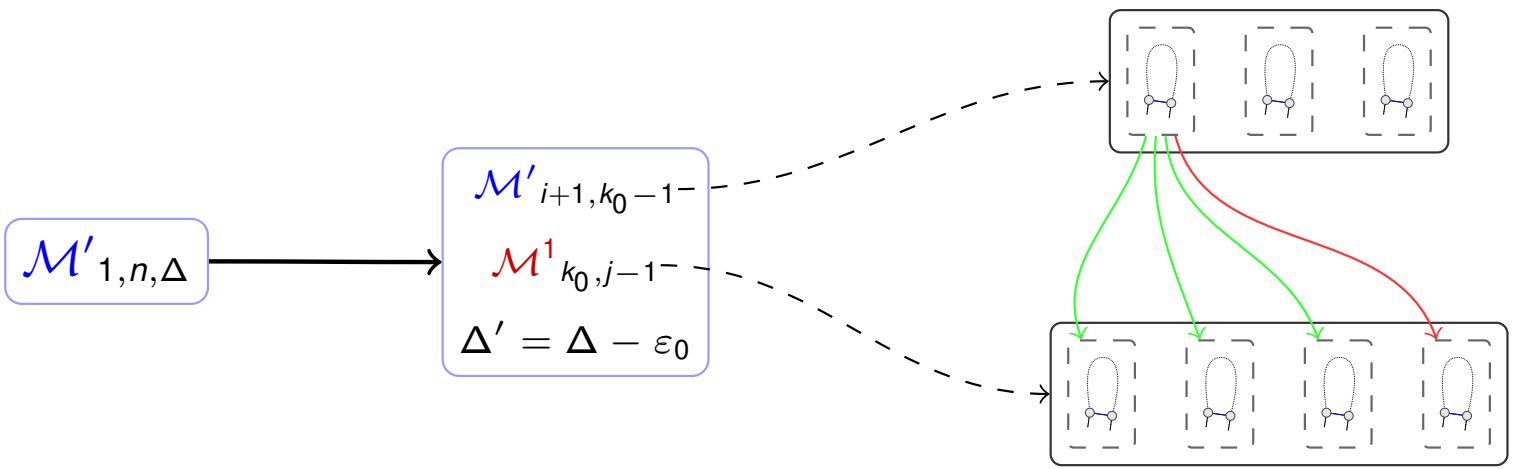
- Compute minimum free-energy matrices
- Backtrack on any contribution within Δ of MFE;
- Update Δ such that future backtracks create ≥ 1 struct.
- Recursively generate subopts and combine (brute-force ou Sort)



Prob.: Simplified energy model (no pseudoknots, only canonical BPs)
$\Rightarrow$ **Native** structure (functional) could be **overthrown**.

$\Rightarrow$ Investigate suboptimal structures (`RNASubopt` [WFHS99]),
i.e. build all structures within Δ KCal.mol^{-1} of MFE:

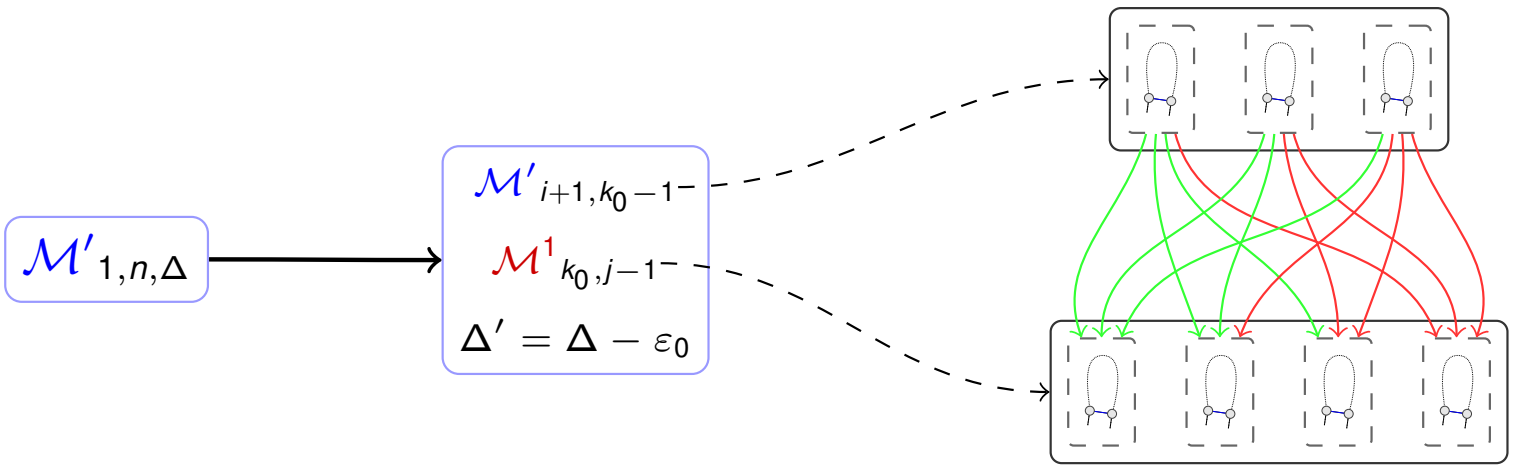
- Compute minimum free-energy matrices
- Backtrack on any contribution within Δ of MFE;
- Update Δ such that future backtracks create ≥ 1 struct.
- Recursively generate subopts and combine (brute-force ou Sort)



Prob.: Simplified energy model (no pseudoknots, only canonical BPs)
$\Rightarrow$ **Native** structure (functional) could be **overthrown**.

$\Rightarrow$ Investigate suboptimal structures (`RNASubopt` [WFHS99]),
i.e. build all structures within Δ KCal.mol^{-1} of MFE:

- Compute minimum free-energy matrices
- Backtrack on any contribution within Δ of MFE;
- Update Δ such that future backtracks create ≥ 1 struct.
- Recursively generate subopts and combine (brute-force ou Sort)



Prob.: Simplified energy model (no pseudoknots, only canonical BPs)
$\Rightarrow$ **Native** structure (functional) could be **overthrown**.

$\Rightarrow$ Investigate suboptimal structures (`RNASubopt` [WFHS99]),
i.e. build all structures within Δ KCal.mol^{-1} of MFE:

- Compute minimum free-energy matrices
- Backtrack on any contribution within Δ of MFE;
- Update Δ such that future backtracks create ≥ 1 struct.
- Recursively generate subopts and combine (brute-force ou Sort)

$$\Rightarrow \text{Time complexity (Sort)} : \mathcal{O}(n^3 + n \cdot k \log(k))$$

(k grows exponentially fast with Δ!)

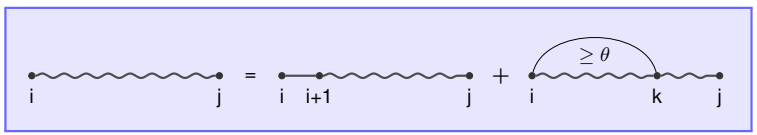
What is a good dynamic programming scheme?

Correction of a (Ensemble) dynamic programming scheme:

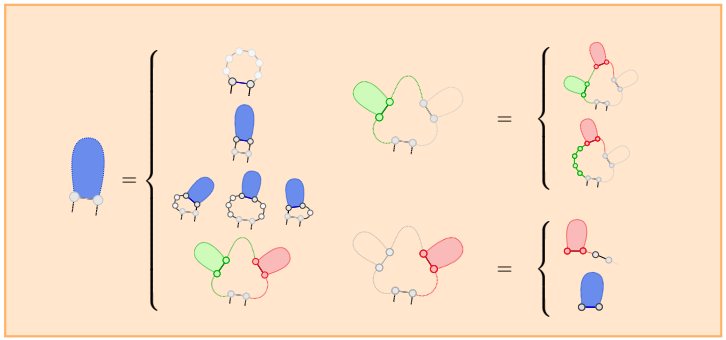
Objective function correctly computed/inherited at local level

\+ All the conformations can be obtained

$\Rightarrow$ Correct algorithm (Induction)



$\Leftrightarrow ?$



Enumerating search space helps **but** does not constitute a proof.

Need to **show equivalence** of DP schemes, *e.g.* use one to simulate the other and vice versa.

(Generating functions may help)

Correction of a (Ensemble) dynamic programming scheme:

Objective function correctly computed/inherited at local level

\+ All the conformations can be obtained

$\Rightarrow$ Correct algorithm (Induction)

$$C_{i,t} = 1, \quad \forall t \in [i, i+\theta]$$

$$C_{i,j} = \sum \left\{ \begin{array}{l} C_{i+1,j} \\ \sum_{k=i+\theta+1}^{j} 1 \times C_{i+1,k-1} \times C_{k+1,j} \end{array} \right.$$

Homopolymère (Toute paire autorisée) + $\theta = 1$
$\Rightarrow C_{1,n} = \mathbf{1}, 1, 1, 2, 4, 8, 17, 32, 82, 185, 423, \ldots$

$\Leftrightarrow$?

$$\mathcal{C'}_{i,j} = \sum \left\{ \begin{array}{l} 1 \\ \mathcal{C'}_{i+1,j-1} \\ \sum_{i',j'} \mathcal{C'}_{i',j'} \\ \sum_{k} \mathcal{C}_{i+1,k-1} \times \mathcal{C}^1{}_{k,j-1} \end{array} \right.$$

$$\mathcal{C}_{i,j} = \sum_{k} \left((\mathcal{C}_{i,k-1} + 1) \times \mathcal{C}^1{}_{k,j} \right)$$

$$\mathcal{C}^1{}_{i,j} = \mathcal{C}^1{}_{i,j-1} + \mathcal{C'}_{i,j}$$

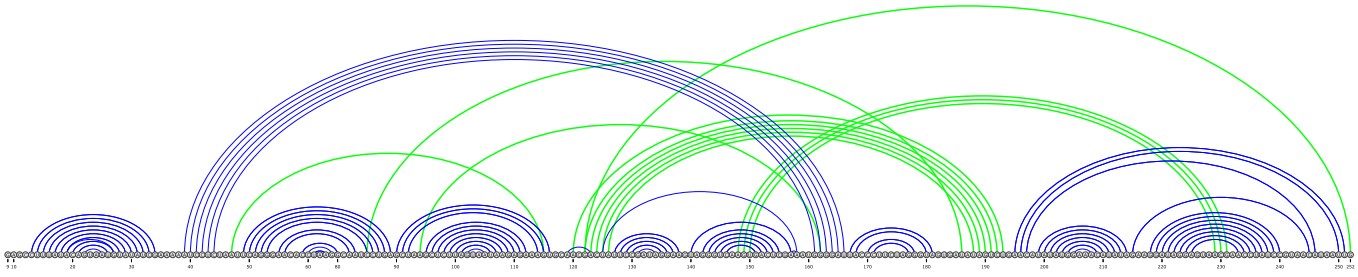
Homopolymère + $\theta = 1$
$\Rightarrow \mathcal{C'}_{1,n} = \mathbf{0}, 1, 1, 2, 4, 8, 17, 32, 82, 185, 423, \ldots$

Enumerating search space helps **but** does not constitute a proof.

Need to **show equivalence** of DP schemes, *e.g.* use one to simulate the other and vice versa.

(Generating functions may help)

Pseudoknots are essential to the folding and activity of multiple RNA families.



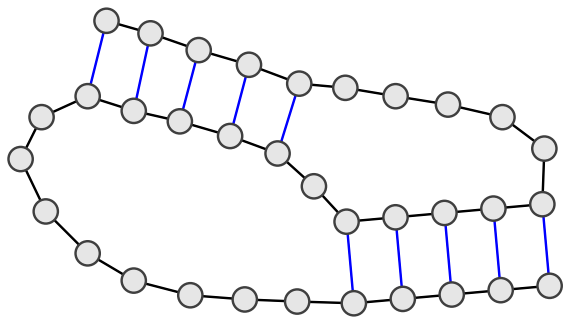
Groupe I Ribozyme

Their disregard within current folding algorithms stems both from **algorithmic** and **energetic** intricacies.

(**Pseudoknots =** Crossings $\Rightarrow$ foldings delimited by base-pair can no longer be assumed to be independent)

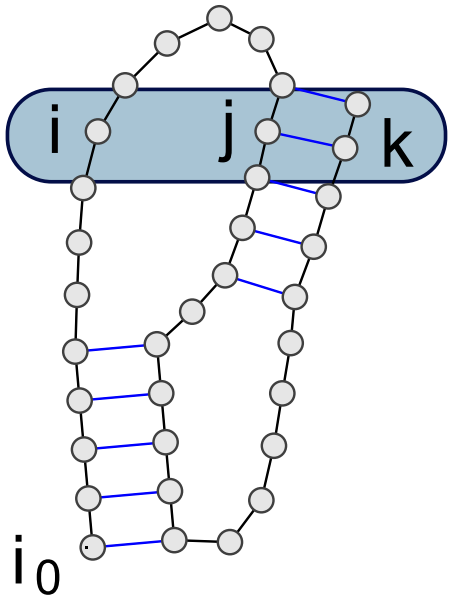
Type	Complexity	Reference
Secondary structures	$\mathcal{O}(n^3)$	[MSZT99]
L&P	$\mathcal{O}(n^5)$	[LP00]
D&P	$\mathcal{O}(n^5)$	[DP03]
A&U	$\mathcal{O}(n^5)$	[Aku00]
R&E	$\mathcal{O}(n^6)$	[RE99]
Unconstrained	NP-complete	[LP00]

Goal: Capture a category of **simple**, **yet** recurrent, pseudoknots.



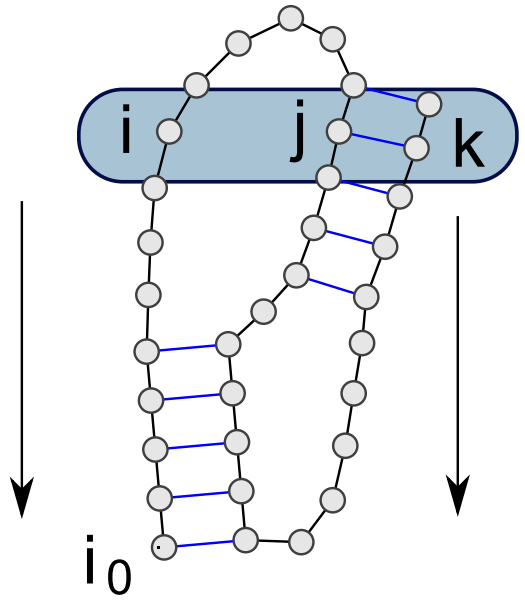
Idea: When such a PK motif is **rotated**, one can deduce the MFE of a triplet (i, j, k) from the MFE of triplets **directly below** it.

Goal: Capture a category of **simple**, **yet** recurrent, pseudoknots.

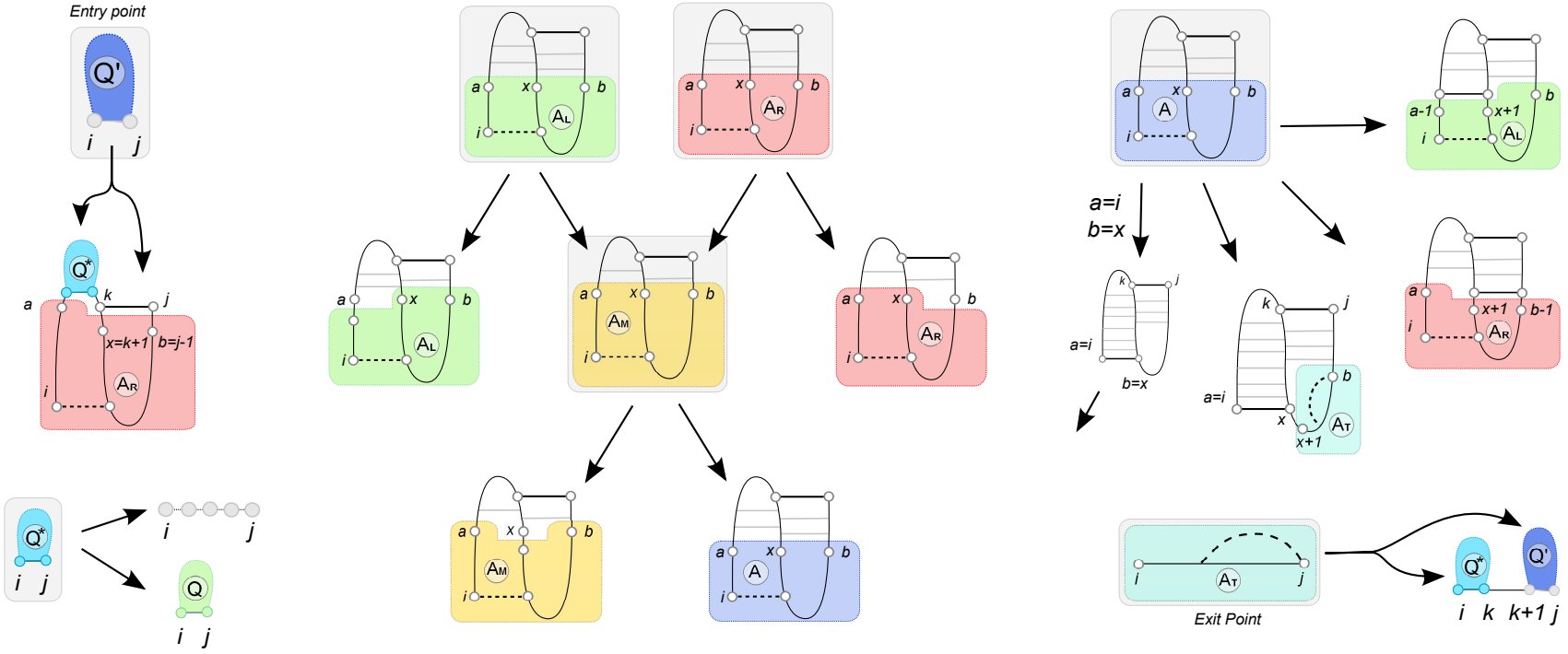


Idea: When such a PK motif is **rotated**, one can deduce the MFE of a triplet (i, j, k) from the MFE of triplets **directly below** it.

Goal: Capture a category of **simple**, **yet** recurrent, pseudoknots.



Idea: When such a PK motif is **rotated**, one can deduce the MFE of a triplet (i, j, k) from the MFE of triplets **directly below** it.



Application/Problem	Weight fun.	Time/Space	Ref.
Energy minimization	π_{bp}	$\mathcal{O}(n^4)/\mathcal{O}(n^4)$	[Aku00]
Partition function	$e^{\frac{-\pi_{bp}}{RT}}$	$\mathcal{O}(n^4)/\mathcal{O}(n^4)$	$\Theta(n^6)$[CC09]
BP probabilities	$e^{\frac{-\pi_{bp}}{RT}}$	$\mathcal{O}(n^4)/\mathcal{O}(n^4)$	–
Sampling (k-struct.)	$e^{\frac{-\pi_{bp}}{RT}}$	$\mathcal{O}(n^4 + kn\log n)/\mathcal{O}(n^4)$	–

Exercice: Write DP equation for MFE computation, counting and partition function.

Hypothesis: Common evolutionary pressure = Common function .

Within certain RNA families (ex.: RNAse-P), low sequence conservation **yet** high structural conservation.

Algorithmic problems:

- **Editing:** Compute *distance* between two secondary structures A and B. Find minimal cost sequence of operations to turn A into B. Already NP-complete for two secondary structures [BFRS07].
- **Alignment:** Find minimal cost super-structure.
  Generalizes sequence alignment. Polynomial ($\mathcal{O}(n^4)$) for secondary structures [BDD+08], NP-complete in 3D [SZS+08].
  Alternatives: Local/global alignment, motifs search (aka small-in-large).
- **Superimposition:** Find solid-body geometric transform (Rotation, translation, zoom) to superimpose *as well as possible* the coordinates of two RNAs having known matching. Polynomial in 3D [McL82].

Remark: Algorithmic hardness stems from finding the matching (i.e. combinatorial, not geometric).

When 3D models are available, the alignment problem can be tackled in a purely geometric setting.

Problem

Input: Motif m, target structure b (ordered set of 3D points).
Output: Matching of m *versus* a subset of b that minimizes a notion of geometric discrepancy.

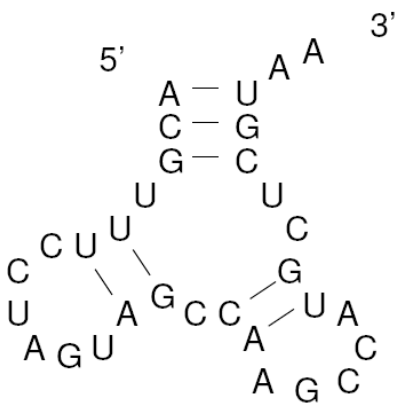
Geometric discrepancy: In FR3D [SZS+08], a **discrepancy** function D combines two error functions L et A, respectively accounting for the superimposability (L) and base orientation (A) of m and b.

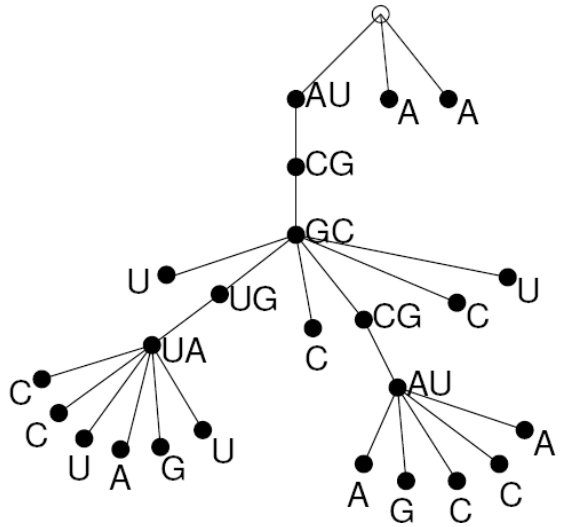
$$L = \sqrt{\min_{R,T} \sum_{i=1}^{m} \|b_i - R(T(m_i))\|^2} \quad A = \sqrt{\sum_{i=1}^{m} \alpha_i^2} \quad D = \frac{1}{m}\sqrt{L^2 + A^2}$$

R, T: Rotation and translation. c_i: Center of mass (CM) of base m_i. α_i: Spread between orientation of CMs/bases in m_i et b_i.

Backtrack + Incremental pruning (Bounds on D) $\Rightarrow$ Combinatorial explosion!
But exact search feasible for smaller motifs.

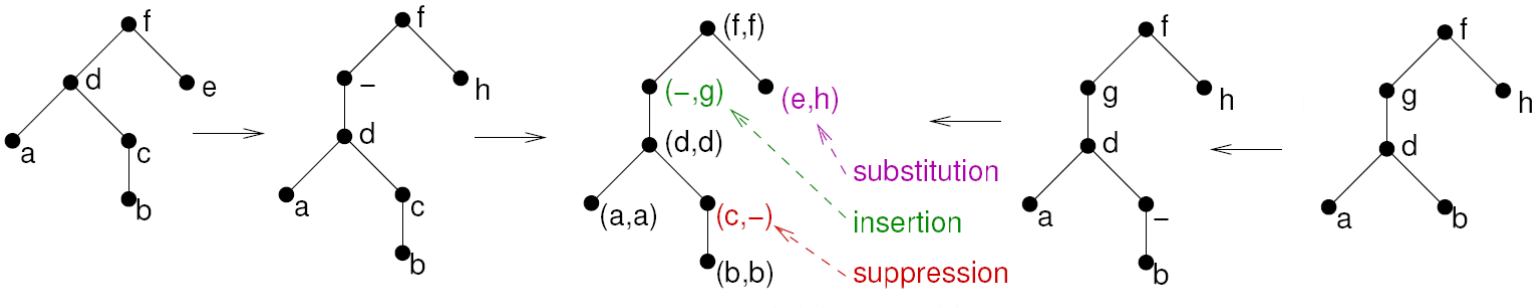
The alignment of two secondary structures is based on their tree-like representations[1].





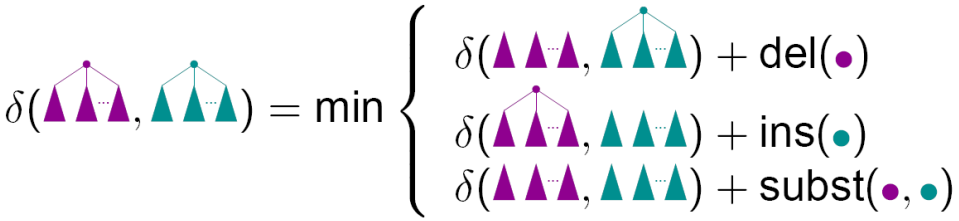
Base pairs ⇒ internal nodes Unpaired bases ⇒ Leaves

Alignment = Complete matching having minimal cost.

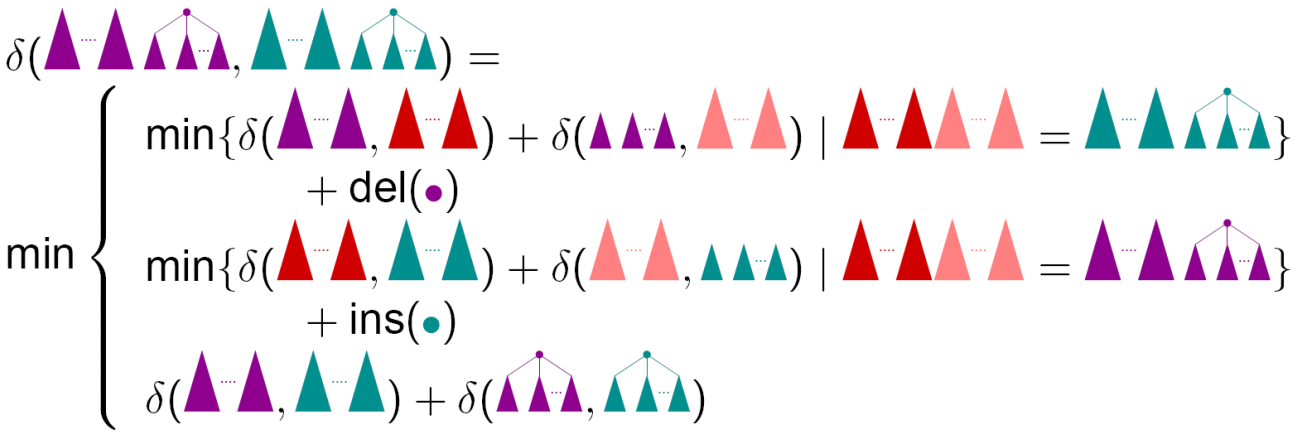


[1] Illustrations empruntées à C. Herrbach

Aligning Trees[2]



Aligning Forests

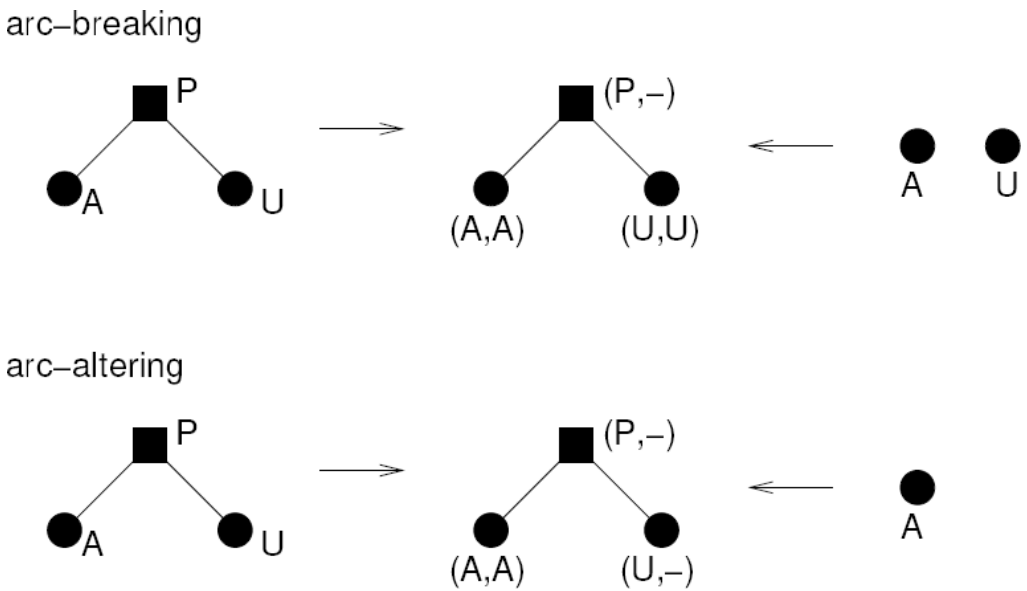


Worst-case complexity in $\mathcal{O}(n^4)$ [JWZ94], on average in $\mathcal{O}(n^2)$ [HDD07].
But RNA-specific operations are lacking

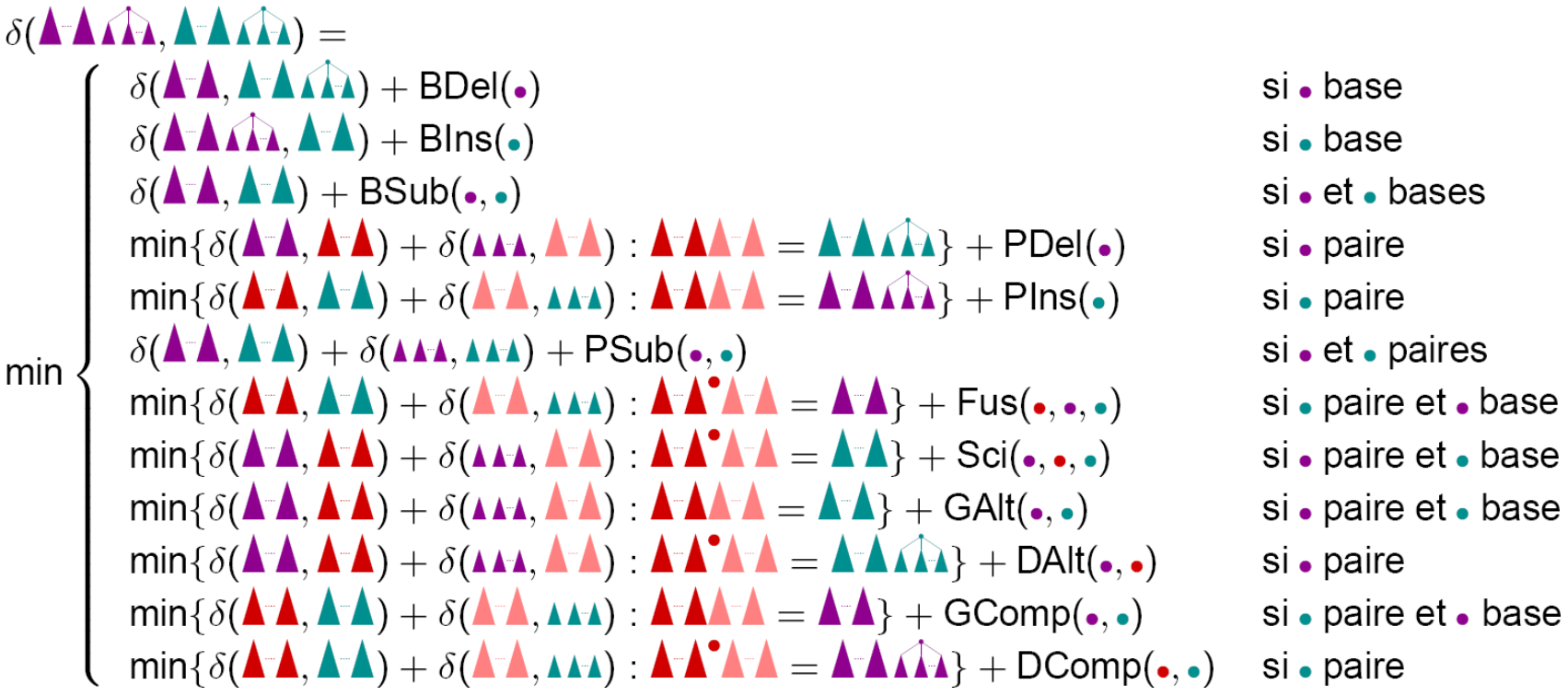
[2] Idem

Parametrization of operation costs, but some operations, atomic in a realistic model, must be composed from available ones.
Example: To detach a base-pair, delete node (base-pair), and insert two leaves (bases).

RNAForester: Based on Jiang, Wang & Zhang algorithm
+ Integration of RNA-specific operations[3].

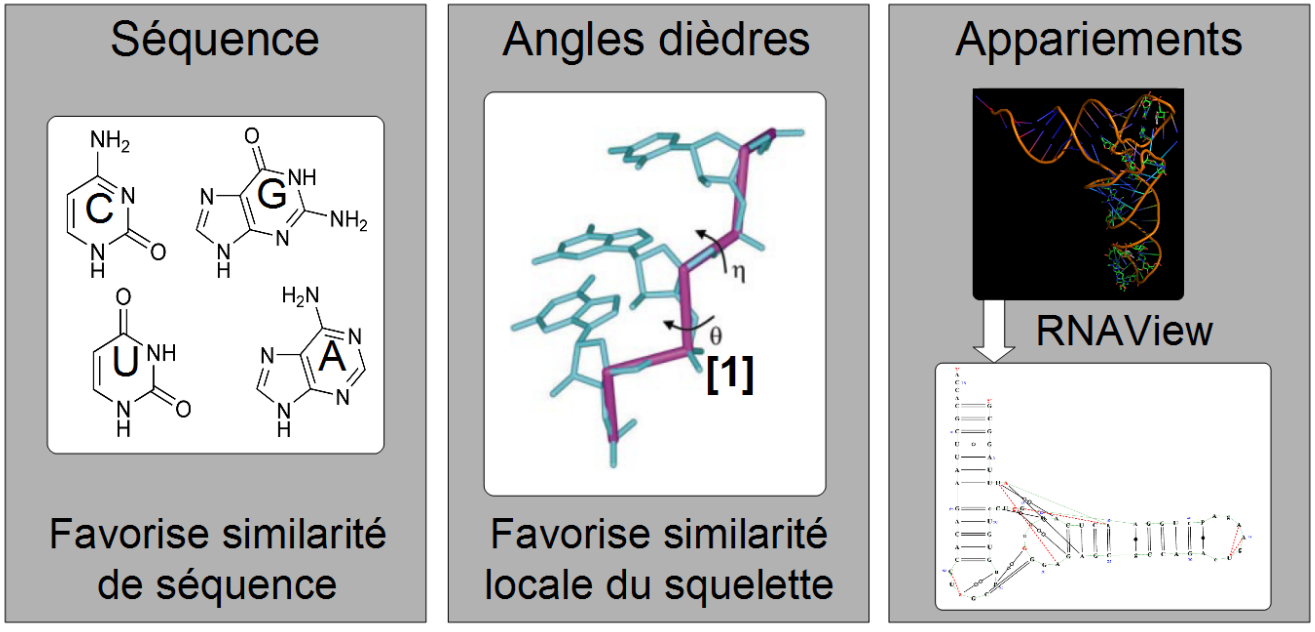


[3] Idem


δ(,) =
min
δ(,) + BDel(•) si • base
δ(,) + BIns(•) si • base
δ(,) + BSub(•, •) si • et • bases
min{δ(,) + δ(,) : = } + PDel(•) si • paire
min{δ(,) + δ(,) : = } + PIns(•) si • paire
δ(,) + δ(,) + PSub(•, •) si • et • paires
min{δ(,) + δ(,) : = } + Fus(•, •, •) si • paire et • base
min{δ(,) + δ(,) : = } + Sci(•, •, •) si • paire et • base
min{δ(,) + δ(,) : = } + GAlt(•, •) si • paire et • base
min{δ(,) + δ(,) : = } + DAlt(•, •) si • paire
min{δ(,) + δ(,) : = } + GComp(•, •) si • paire et • base
min{δ(,) + δ(,) : = } + DComp(•, •) si • paire

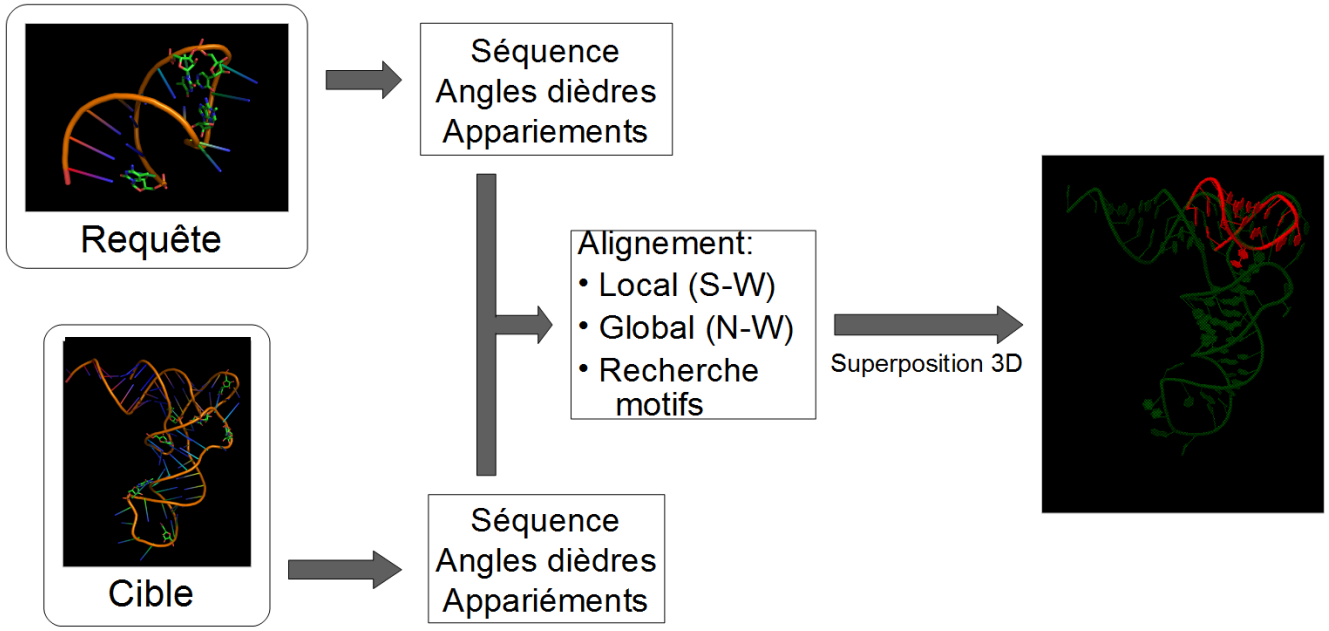
DIAL [FPLC07] is an integrative method which focuses on local similarities.
Idea: RNA is flexible, meaningless local variations (even of small amplitudes) may induce large geometric discrepancies.

DIAL captures local similarities at three levels:



DIAL [FPLC07] is an integrative method which focuses on local similarities.
Idea: RNA is flexible, meaningless local variations (even of small amplitudes) may induce large geometric discrepancies.

A sequence alignment algorithm is then used




Tatsuya Akutsu.
Dynamic programming algorithms for rna secondary structure prediction with pseudoknots.
Discrete Appl. Math., 104(1-3):45–62, 2000.


G. Blin, A. Denise, S. Dulucq, C. Herrbach, and H. Touzet.
Alignment of rna structures.
Transactions on Computational Biology and Bioinformatics, ...:..., 2008.
A paraître.


Guillaume Blin, Guillaume Fertin, Irena Rusu, and Christine Sinoquet.
Extending the Hardness of RNA Secondary Structure Comparison.
In Bo Chen, Mike Paterson, and Guochuan Zhang, editors, *ESCAPE'07*, volume 4614 of *LNCS*, pages 140–151, Hangzhou, China, Apr 2007.


S. Cao and S-J Chen.
Predicting structured and stabilities for h-type pseudoknots with interhelix loop.
RNA, 15:696–706, 2009.


Robert M Dirks and Niles A Pierce.
A partition function algorithm for nucleic acid secondary structure including pseudoknots.
J Comput Chem, 24(13):1664–1677, Oct 2003.


F. Ferrè, Y. Ponty, W. A. Lorenz, and Peter Clote.
Dial: A web server for the pairwise alignment of two RNA 3-dimensional structures using nucleotide, dihedral angle and base pairing similarities.
Nucleic Acids Research, 35(Web server issue):W659–668, July 2007.


Claire Herrbach, Alain Denise, and Serge Dulucq.
Average complexity of the jiang-wang-zhang pairwise tree alignment algorithm and of a rna secondary structure alignment algorithm.
In *Proceedings of MACIS 2007, Second International Conference on Mathematical Aspects of Computer and Information Sciences*, 2007.


M. Hochsmann, B. Voss, and R. Giegerich.
Pure multiple RNA secondary structure alignments: A progressive profile approach.
01(1):53–62, 2004.


Tao Jiang, Lusheng Wang, and Kaizhong Zhang.
Alignment of trees - an alternative to tree edit.
In *CPM '94: Proceedings of the 5th Annual Symposium on Combinatorial Pattern Matching*, pages 75–86, London, UK, 1994. Springer-Verlag.


R. B. Lyngsøand C. N. S. Pedersen.
RNA pseudoknot prediction in energy-based models.
Journal of Computational Biology, 7(3-4):409–427, 2000.


D. McLachlan.
Rapid comparison of protein structures.
Acta cristallographica A, 38(6):871–873, 1982.


D. H. Mathews, J. Sabina, M. Zuker, and D. H. Turner.
Expanded sequence dependence of thermodynamic parameters improves prediction of rna secondary structure.
Journal of Molecular Biology, 288(5):911–940, May 1999.


E. Rivas and S.R. Eddy.
A dynamic programming algorithm for RNA structure prediction including pseudoknots.
J Mol Biol, 285:2053–2068, 1999.


M. Sarver, C. Zirbel, J. Stombaugh, A. Mokdad, and N. B. Leontis.
FR3D: Finding local and composite recurrent structural motifs in RNA 3D.
Journal of Mathematical Biology, 56(1–2):215–252, January 2008.


S. Wuchty, W. Fontana, I.L. Hofacker, and P. Schuster.
Complete suboptimal folding of RNA and the stability of secondary structures.
Biopolymers, 49:145–164, 1999.